Topics in Molecular and Structural Biology

OLIGODEOXYNUCLEOTIDES

Antisense Inhibitors of Gene Expression

Edited by

Jack S. Cohen

National Cancer Institute
National Institutes of Health
Bethesda, Maryland

CRC Press, Inc.
Boca Raton, Florida

7

Control of Gene Expression by Oligodeoxynucleotides Covalently Linked to Intercalating Agents and Nucleic Acid-cleaving Reagents

Claude Hélène and Jean-Jacques Toulmé

1 Introduction

Gene expression in all living organisms is controlled at different steps of information processing: transcription of DNA into premessenger RNAs; splicing of mRNA precursors; post-transcriptional modifications of mRNAs (capping, polyadenylation); transfer of mRNAs from the nucleus to the cytoplasm; translation of mRNA; mRNA stability .... In most cases this regulation is achieved by proteins that bind to specific regions of DNA or RNA and either block or stimulate the enzymatic processes (see Hélène and Lancelot, 1982, for a review). Recently it has been shown that small RNAs could play a role similar to that of regulatory proteins. Upon hybridization with a messenger RNA, these regulatory RNAs may alter the translation process or induce premature termination of transcription (see Green *et al.*, 1986, for a review). These regulatory processes have been originally observed in bacteria (Green *et al.*, 1986) but they might also occur in eukaryotes (Heywood, 1986). The discovery of regulatory RNAs has been the starting point for the design of 'antisense' RNAs. By inserting a gene fragment close to a strong promoter in the reverse orientation as compared with that of the gene itself, the non-template strand of the gene fragment is now used as a template by RNA polymerase. As a consequence this 'antisense' transcript is fully complementary to the mRNA. This might block mRNA translation or other post-transcriptional processes such as splicing or mRNA migration from the nucleus to the cytoplasm (Kim and Wold, 1985; Green *et al.*, 1986).

The idea of using synthetic oligonucleotides complementary to RNA sequences to alter gene expression was put forward in several laboratories (Paterson *et al.*, 1977; Stephenson and Zamecnik, 1978; Summerton, 1979; Jayaraman *et al.*, 1981; Trudel *et al.*, 1981; Asseline *et al.*, 1983, 1984a, b; for reviews see also Knorre and Vlassov, 1985; Hélène *et al.*, 1985; Hélène, 1987; Stein and Cohen, 1988; Toulmé and Hélène, 1988). It has been shown that oligodeoxynucleotides complementary to mRNAs could block translation in acellular systems, in microinjected *Xenopus* oocytes and in cells in culture. There is an obvious need for developing new families of gene regulatory substances that could be used *in vivo* to control the expression of undesirable genes, such as oncogenes, or to inhibit the development of viruses or parasites. The application of oligodeoxynucleotides to *in vivo* studies faces two main problems: (1) their penetration into living cells in culture is limited; (2) their sensitivity to nucleases makes their lifetime very short (Cazenave *et al.*, 1987b).

Several attempts have been made to overcome these two difficulties. The phosphodiester backbone of the oligodeoxynucleotide can be changed to a methylphosphonate backbone; the loss of negative charges makes these oligophosphonates more efficient in penetrating through the cell membranes and much more resistant to nucleases (Miller *et al.*, 1983). The phosphate group can be replaced by a phosphorothioate (Marcus-Sekura *et al.*, 1987); these oligophosphorothioates are much more resistant to nucleases than natural oligonucleotides. Attachment of oligonucleotides to polymers such as poly-L-lysine increases the efficiency of penetration and makes oligonucleotides active at much lower concentrations (Lemaître *et al.*, 1987).

This review summarizes the approach we have been following to design new families of specific gene regulatory substances. A nucleic acid base sequence can be easily recognized by an oligonucleotide of complementary sequence. The stability of the mini-double helix formed by an oligonucleotide with its target sequence can be increased by covalent attachment of an intercalating agent at one end of the oligonucleotide (Asseline *et al.*, 1983, 1984a,b). In addition, the intercalating agent endows the oligonucleotide with a higher penetration across cell membranes and stabilizes it against 3'- or 5'-exonucleases, depending on the attachment site (Verspieren *et al.*, 1987). The other end of the oligonucleotide can be substituted by a reagent which can be activated to modify the target sequence by either chemical or photochemical activation (Boidot-Forget *et al.*, 1986; Le Doan *et al.*, 1987a; Praseuth *et al.*, 1987, 1988a). Specific cleavage of a mRNA target or chemical modification of the bases at the binding site of the oligonucleotide should prevent translation of the mRNA. In addition, the oligonucleotide can be modified in such a way as to make it more resistant to nucleases, e.g. by substituting synthetic α-anomers of nucleotides for the natural β-anomers.

Oligodeoxynucleotides can recognize not only mRNAs but also duplex DNA by binding to the major groove. Therefore oligonucleotides can be

used to control gene expression at the transcriptional level. Oligonucleotides carrying a reactive group can induce irreversible reactions in duplex DNA, including double-strand cleavage.

2 Oligodeoxynucleotides as Anti-messengers

Specificity of Oligonucleotide Targeting to Unique Sequences

Targeting to Genomic DNA

The minimum size that an oligonucleotide should have in order to recognize a single specific sequence in a genome can be calculated on the basis of different assumptions.

Assuming a statistical distribution of base pairs in a genome characterized by a fraction f of A.T base pairs ($f = [A.T]/[A.T] + [G.C]$), the probability (p_0) of finding a sequence of n nucleotides is given by Equation (1):

$$p_0 = [(f/2)]^{(a+t)} \times [(1-f)/2]^{(g+c)} \quad (1)$$

where a, t, g, c are the numbers of adenines, thymines, guanines and cytosines in the oligonucleotide ($n = a + t + g + c$). The number (Q) of identical sequences of n nucleotides in a genome containing N base pairs is given by Equation (2):

$$Q = p_0 \times 2N \quad (2)$$

where the factor 2 accounts for the presence of the two strands in DNA. For *E. coli*, which contains about 4.5×10^6 base pairs in its genome with equal numbers of A.T and G.C base pairs ($f = 0.5$), the minimal length (n) that an oligonucleotide should have in order to find a single complementary sequence ($Q \leq 1$) is $n = 12$. In human cells, with $N = 4 \times 10^9$ and $f \approx 0.6$, the minimal length is calculated to vary from $n = 15$ if the oligonucleotide contains only Gs and Cs to $n = 19$ if the oligonucleotide contains only As and Ts.

The above calculation assumes a statistical distribution of base pairs in the genome, which, of course, is not correct. Analyses of nearest-neighbour frequencies have shown that the dinucleotide CpG is underrepresented in eukaryotic genomes. The probability of finding an oligonucleotide sequence in a genome can be calculated on the basis of nearest-neighbour frequencies using a first-order Markov chain. According to Markov's theory, this probability is equal to the product of probabilities of all overlapping dinucleotides in the sequence divided by the product of probabilities of shared mononucleotides. For example, the probability of finding the sequence CATCGT is given by Equation (3):

$$p_1(\mathrm{CATCGT}) = \frac{p(\mathrm{CA}) \times p(\mathrm{AT}) \times p(\mathrm{TC}) \times p(\mathrm{CG}) \times p(\mathrm{GT})}{p(\mathrm{A}) \times p(\mathrm{T}) \times p(\mathrm{C}) \times p(\mathrm{G})} \quad (3)$$

where $p(XY)$ is the probability of finding the sequence XpY, and $p(X)$ is the probability of finding nucleotide X.

A second-order Markov chain can be used if the probability of finding trinucleotides ($p(XYZ)$) is known, as shown in Equation (4):

$$p_2(CATCGT) = \frac{p(CAT) \times p(ATC) \times p(TCG) \times p(CGT)}{p(AT) \times p(TC) \times p(CG)} \quad (4)$$

Higher-order Markov chains can be used if the frequency of longer sequences is known (tetranucleotides, pentanucleotides . . .). With the accumulation of sequence data, such calculations should become more and more accurate in predicting the probability of finding an oligonucleotide sequence in unknown regions of the genome.

Table 7.1 gives the probability of finding the decanucleotide sequence GGCATCGTCG in the *E. coli* and human genomes, according to equations (1) and (3). This sequence is found in the murine and human c-*myc* genes and was used in our laboratory to inhibit *myc* mRNA *in vitro* translation (see below). In *E. coli* the calculated probability does not markedly change with the mode of calculation. Obviously this is not the case for the human genome, because the chosen sequence contains two CpG dinucleotides which are underrepresented as compared with a statistical distribution. Equation (3) should be preferred to calculate the probability of finding any oligonucleotide sequence in eukaryotic genomes. Since we are interested in using oligonucleotides to specifically regulate gene expression, the probability that should be calculated is not that of finding the oligonucleotide but rather that of finding the complementary sequence. These two probabilities are obviously equal if the genome is the target, owing to the complementarity of the two strands in the DNA double helix. This is no longer true if the oligonucleotide is targeted to a messenger RNA. However, the only dinucleotide that significantly appears less frequently than calculated on a statistical basis is CpG in eukaryotes. Its complementary sequence is also CpG. Therefore, the calculated probabilities will not be very different if an oligonucleotide or its complementary sequence is considered.

Table 7.1 Calculated probability of finding the sequence GGCATCGTCG in *E. coli* and human genomes calculated on the basis of zero-order and first-order Markov chains (equations 1 and 3, respectively)

	E. coli	*Human*
p_0[a]	9.5×10^{-7}	4.1×10^{-7}
p_1[b]	11×10^{-7}	0.22×10^{-7}

[a] Equation (1).
[b] Equation (3).
The frequency of nucleotides and dinucleotides was obtained from the *Handbook of Biochemistry and Molecular Biology.*

Targeting to Messenger RNAs

Only a small fraction of the eukaryotic genome is transcribed into messenger RNA in a living cell at a given time. Only one strand of the DNA is usually transcribed, even though there is increasing evidence that divergent transcription of both DNA strands takes place in eukaryotes. It has been estimated that about 0.5% of the genomic DNA is transcribed into mRNA. Therefore, the probability of finding an oligonucleotide sequence is much lower in the mRNA population than in DNA, since the target size drops from 4×10^9 to 2×10^7 units in human cells. On this assumption the minimal length that an oligonucleotide should have in order to find a single target at the mRNA level is reduced: from 15 to 11 if the oligonucleotide contains only Gs and Cs and from 19 to 15 if it contains only As and Ts. Owing to the low frequency of CpG dinucleotides in the human genome (see above), the length of an oligonucleotide targeted to a sequence containing several CpG dinucleotides might be chosen quite short without losing the specificity of mRNA recognition. The results presented in Table 7.1 show that the decanucleotide sequence GGCATCGTCG should not occur more than once, on a statistical basis, in the human mRNA population (the calculated probability is 0.22×10^{-7} and the target mRNA complexity amounts to 2×10^7 nucleotides; see above).

When a viral RNA or a viral mRNA is chosen as a target for oligonucleotides, it should be kept in mind that codon usage is usually different from that of the host cell. There might be a difference in trinucleotide and tetranucleotide frequencies between the virus and its host. This difference can be taken into account when choosing a target sequence.

The conclusion that can be drawn from the above considerations is that a high specificity of oligonucleotide binding to mRNAs can be achieved with quite short oligonucleotides. Using short oligonucleotides provides additional advantages.

(1) The probability of forming intramolecular secondary structures (hairpins) is lower in short oligonucleotides.
(2) The probability of finding the target sequence in an accessible region of a messenger RNA or a viral RNA is increased. It should be remembered that RNAs adopt folded conformations. Complementary sequences which can be far apart in the primary sequence may form duplex structures. Hairpins may fold on themselves to engage in tertiary interactions (as observed, e.g., in the folded clover-leaf structure of tRNAs).
(3) The penetration of negatively charged oligonucleotides across cell membranes is expected to decrease when the length of the oligonucleotide increases. Therefore, short oligonucleotides could penetrate better inside living cells.
(4) If oligonucleotides have to be chemically modified in order to make them, e.g., more resistant to nucleases, chemical synthesis and purifi-

cation should be easier to achieve for short molecules (also they should cost less!). These points might be of special interest if practical (therapeutical) applications are contemplated.

(5) The specificity of interaction will be higher with shorter oligonucleotides provided that the target (complementary) sequence is found only once. Studies on inhibition of gene expression in living cells are carried out under well-defined conditions of temperature, ionic concentrations, etc., which are imposed upon the experimentalist by the species under investigation. For example, experiments on human cells are performed at 37 °C. The important parameters which should be considered in choosing oligonucleotide length are the stability (the free energy) of binding to the target sequence *at 37 °C* and the discrimination between closely related sequences at this temperature (which determines the specificity of inhibition). The free energy cost due to a mismatch does not change appreciably when an oligonucleotide is elongated, except when the mismatch is located close to the oligonucleotide end which is elongated. But its relative contribution to complex destabilization of course decreases when the oligonucleotide length increases.

(6) An important component in the biological activity of oligodeoxynucleotides is RNase-H, an enzyme which cuts RNA within the region hybridized to the oligodeoxynucleotide (see below). This enzyme recognizes duplexes as short as 4 base pairs (Donis-Keller, 1979) and could be responsible for non-specific inhibition of gene expression due to partial complementarity. Such an effect should be minimized by using short oligonucleotides.

Oligonucleotides Covalently Linked to Intercalating Agent

Increased Binding to Target Sequences

The above calculations suggest that short synthetic oligonucleotides can be designed to bind to a single nucleic acid target. However, a short oligonucleotide might not have a strong enough affinity towards its target sequence if the number of base pairs involved is too small. There are different ways of increasing this affinity. We have chosen to covalently link an intercalating agent to one (or both) oligonucleotide end(s). Intercalating agents are polycyclic aromatic molecules that insert their planar aromatic ring between two consecutive base pairs of double-stranded DNA. They bind much more weakly to single-stranded structures. If the linker between the oligonucleotide and the intercalating agent is appropriately chosen, intercalation can occur in the mini-double helix formed when the oligonucleotide is bound to its complementary sequence (Figure 7.1). To a first approximation the free energy of binding of the composite molecule (ONBI, for *O*ligo*N*ucleotide-*B*ridge-*I*ntercalator) should be the sum of the free energy for binding the oligonucleotide to its complementary sequence (ΔG_{ON}) and that for intercalation (ΔG_I), corrected for an entropy term ($T\Delta S_m$) taking into account the restricted configurational space available to the intercalating agent when it is covalently linked to the oligonucleotide:

$$\Delta G_{ONBI} = \Delta G_{ON} + \Delta G_I - T\Delta S_m \quad (5)$$

Since ΔS_m in Equation (5) is positive, the association constant for the ONBI ($K_{ONBI} = \exp - \Delta G_{ONBI}/RT$) should be at least the product of the association constants for the oligonucleotide and the intercalating agent.

$$K_{ONBI} = \alpha K_{ON} \times K_I \quad (6)$$

with $\alpha = \exp(\Delta S_m/R) > 1$.

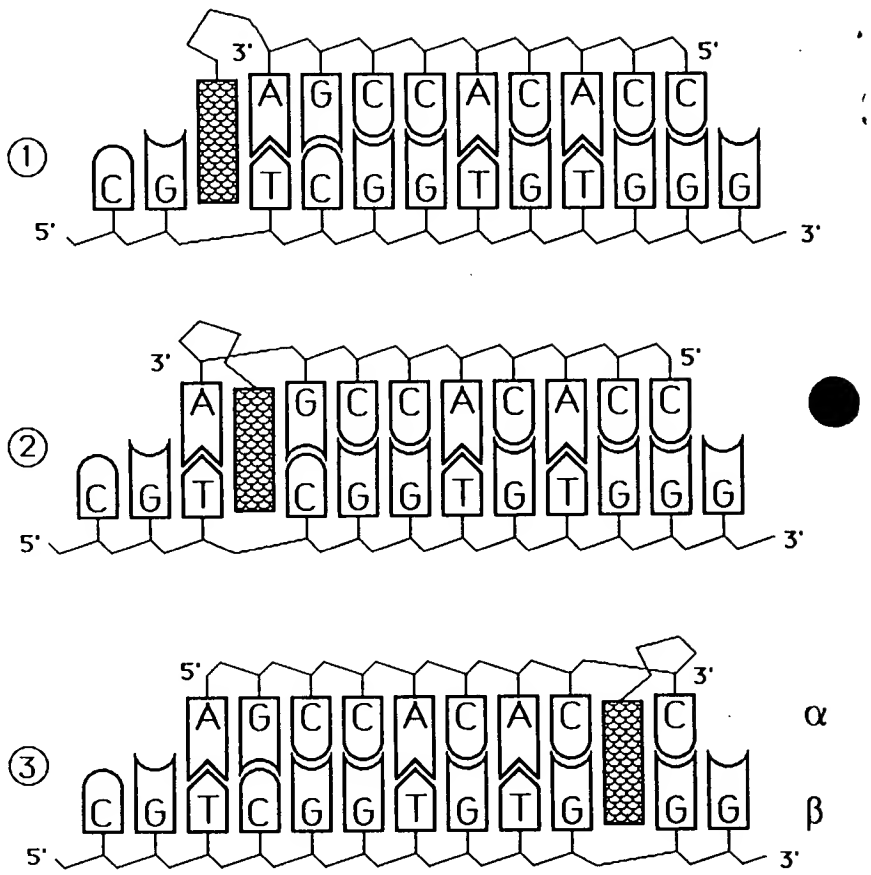

Figure 7.1 Schematic representation of an oligonucleotide covalently linked to an intercalating agent via its 3' end (rectangle) when bound to a complementary sequence. In 1 and 2 an oligonucleotide with natural β-anomers of the nucleotides is depicted with two different positions of the intercalating agent. In 3 an oligonucleotide with synthetic α-anomers is shown bound in a *parallel* orientation with respect to its complementary sequence